Detecting Fraudulent Campaigns on a Social Media Platform with Text-Based Machine Learning

Fernando Ferreira
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
fernando.ferreira@netlab.eco.ufrj.br

Thiago Ciodaro
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
thiago.ciodaro@netlab.eco.ufrj.br

Felipe Fink Grael
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
felipe.grael@netlab.eco.ufrj.br

Vitor do Carmo
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
vitor.carmo@netlab.eco.ufrj.br

Danielle de Pinho Mello
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
danielle.mello@netlab.eco.ufrj.br

Alékis de Carvalho Moreira
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
alekis.moreira@netlab.eco.ufrj.br

João Gabriel Haddad
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
joao.haddad@netlab.eco.ufrj.br

Bruno Mauricio Martins
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
bruno.martins@netlab.eco.ufrj.br

Débora Gomes Salles
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
debora.salles@netlab.eco.ufrj.br

Rose Marie Santini
NetLab - School of Communication
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
marie.santini@eco.ufrj.br

***Abstract*—The growing prevalence of fraudulent digital ads, which frequently exploit institutional imagery and microtargeting strategies, is compounded by constantly evolving patterns designed to evade detection. This scenario underscores the need for machine learning approaches based on natural language processing to identify and classify this content effectively. This study proposes a methodological framework for detecting and analyzing fraudulent advertisements on digital platforms, relying exclusively on textual content. A Support Vector Machine classifier was developed using TF-IDF features and trained on a manually annotated dataset. The model achieved a mean F2 score of 0.93 under nested cross-validation, prioritizing recall to reduce the risk of false negatives in a human-in-the-loop annotation workflow. The model was applied at acquisition time to more than 40,000 ads which 77% were confirmed as fraudulent. These campaigns often impersonated public figures and exploited themes such as health, finance, and public programs. A post hoc interpretability analysis was performed to identify the linguistic markers most indicative of fraud.**

***Index Terms*—Fraud Detection, SVM, Digital Advertising, Natural Language Processing, Data Science Application**

I. Introduction

The proliferation of fraud in digital advertising environments has far-reaching implications for public trust and institutional credibility. Fraudulent content, especially when it masquerades as legitimate communication about health, financial services, or public policy, undermines the sense of security of users through social engineering [1]–[5]. These scams often appropriate the imagery and language of trusted institutions, such as media conglomerates, well-known companies, and celebrities, to appear credible [2], [5]–[7].

Beyond individual harm, these schemes constitute a source of funding for criminal networks. In addition to the creation of underground illegal markets for cybercrime and cyber-related crime, digital technologies facilitate the migration of traditional organized crime online and provide a number of opportunities for fraud, corruption, tax evasion, and other criminal activities.

Digital platforms serve as highly effective environments to target vulnerable populations [2], [5], [6], [8]. Through advanced profiling and microtargeting, these platforms deliver tailored messages that can mislead users into sharing personal information, engaging in malicious content, or purchasing counterfeit and illegal products. Among these platforms, Meta (the parent company of Facebook, Instagram, and Messenger) occupies a central role, both as an enabler and a beneficiary of digital advertising practices [5], [6].

Despite public commitments to transparency, Meta has shown a reluctance to actively combat fraudulent advertisers [9]. Internal reports and journalistic investigations suggest that the company deprioritized moderation measures in favor of profit maximization [9]. According to these reports, Meta registered a 22% increase in ad revenue in 2024, reaching over USD 160 billion, while simultaneously reducing effective control over repeat fraud actors, even when clear patterns of abuse were evident [9].

In this opaque ecosystem, the application of machine learning models that leverage natural language processing (NLP) emerges as a promising countermeasure. Recent studies have explored the use of supervised learning techniques for detecting fraudulent behavior in digital advertisements, including the use of Gradient Boosting, Random Forests, and Support Vector Machines (SVMs) to classify malicious content based on behavioral or textual features [6], [6], [10]–[14]. NLP-based classification, in particular, has demonstrated robust performance in extracting semantic patterns from short and noisy ad texts, enhancing the ability to flag scam content with limited metadata [6], [13], [14].

This study aims to establish a methodological framework for investigating the behavior of fraudulent advertisements on digital platforms, with a focus on campaigns that exploit the image and authority of traditional media companies. Combining expertise from both computational and communication sciences, we conducted a longitudinal analysis of the fraud campaigns detected by our classifier, examining their rhetorical structure and narrative techniques.

The investigation is grounded in the development of a supervised classifier and the application of the SHAP interpretability framework to identify linguistic markers most indicative of fraud. The classifier was implemented using SVM, selected for its good performance and compatibility with SHAP analysis. Model development followed best practices, and a robust nested cross-validation protocol was employed to ensure generalization and minimize overfitting.

The longitudinal analysis was based on a newly acquired dataset of 41,521 advertisements collected between July 1st and November 27th, 2024. This dataset, filtered and prioritized by the classifier, served as the empirical foundation for exploring the structure, timing, and thematic diversity of fraud campaigns involving impersonation of media-affiliated entities.

This paper is structured as follows: Section II introduces the methodological foundations that support our approach. Section III details the construction of the datasets used for both training and analysis, including data acquisition strategies and textual feature representation. Section IV outlines the classifier development process, including the validation protocol and interpretability methods. Section V presents the performance metrics obtained through cross-validation. Section VI provides an exploratory and operational analysis of fraud campaigns, based on five months of classified advertisements. Finally, Section VII summarizes our findings and outlines directions for future research.

II. Methodological Framework

This section presents the methodological foundations used in the construction of the fraud detection system.

A. Data Acquisition

Reliable fraud detection in digital advertising demands access to structured and consistent data. However, digital platforms such as Meta provide limited support for systematic investigations (specially in the Global South), where transparency mechanisms are restricted and inconsistently applied [5].

In our context, ad data was obtained through a combination of sources, highlighting the heterogeneity and fragmentation inherent to Meta's transparency infrastructure:

Meta Ads API
: Offers access exclusively to ads labeled as "sensitive", i.e., content related to politics, elections, and social issues. These ads include metadata such as impressions, targeting criteria, and advertiser identity, and are retained for up to seven years.

Manual Collection
: Initially performed to identify and annotate examples of non-sensitive fraudulent ads, particularly those not retained by the platform due to lack of classification or removal.

Web Interface Scraping
: Subsequently developed tools to systematically capture active advertisements directly from the Meta Ad Library interface. While this approach enabled a broader data collection, it remains limited by the lack of metadata, absence of historical archiving and restricted filtering capabilities (making it suitable primarily for short-window monitoring).

The resulting dataset is therefore *heterogeneous*, composed of entries from multiple collection strategies with varying levels of granularity and permanence.

B. Text Preprocessing and Feature Extraction

To prepare the textual data for NLP analysis, several preprocessing steps are needed [18]. Standard techniques include:

- Lowercasing and punctuation removal to reduce variability;
- Stopword filtering to eliminate high-frequency, low-information words;
- URL removal to reduce noise from external references;
- Emoji tokenization to preserve symbolic cues (e.g., ✓ → "checkmarkbutton").

Once normalized, the text is represented using TF-IDF, which emphasizes terms that are frequent within a document but rare across the corpus. The TF-IDF weight

for a term t in a document d from a corpus D is defined as:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \cdot \log\left(\frac{N}{\text{df}(t)}\right)$$

where $\text{tf}(t, d)$ is the term frequency of t in document d, $\text{df}(t)$ is the number of documents in which t appears and N is the total number of documents in the corpus.

TF-IDF vs. Text Embeddings: While modern embedding techniques (e.g., Word2Vec, BERT) capture semantic similarity, they may obscure crucial lexical patterns needed to identify fraudulent content, specifically in domains where deception depends on specific phrasing or branding imitation. TF-IDF, in contrast, preserves token identity, supports interpretability, and performs reliably on medium-scale datasets with limited computational overhead.

C. Support Vector Machines

SVMs are a class of supervised learning algorithms designed to find an optimal hyperplane that separates data points of different classes with the maximum possible margin. Given a labeled training set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ are the input vectors and $y_i \in \{-1, +1\}$ are the corresponding class labels, the primal form of the linear SVM optimization problem can be expressed as:

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \quad y_i(\mathbf{w}\cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n$$

Here, $\mathbf{w}$ is the normal vector to the hyperplane, b is the bias term, ξ_i are slack variables that allow for soft margin classification, and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error [15].

To address non-linearly separable problems, SVMs employ the *kernel trick*, which consists of implicitly mapping the input vectors into a higher-dimensional feature space where linear separation may become feasible. One of the most widely used kernels is the RBF [16], defined as:

$$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$$

where $\gamma > 0$ defines the influence of individual training examples and controls the smoothness of the resulting decision boundary.

This transformation is particularly beneficial in the context of text classification [17]. Text data, when vectorized using techniques such as TF-IDF, typically result in high-dimensional but sparse feature representations, where each dimension corresponds to a unique token or term. Although the original feature space may already be large, the discriminative patterns necessary to separate classes, such as fraudulent versus legitimate ads, may not be linearly separable within it.

D. Explainable Machine Learning with SHAP

Interpretability techniques are used to understand how machine learning models make predictions [20]. In classification tasks, this includes identifying which input features contribute most to a given output. SHAP (SHapley Additive exPlanations) is one such technique. It is based on Shapley values from cooperative game theory and assigns a numerical value to each feature representing its contribution to the model's prediction [21].

SHAP models explanations as additive feature attributions. It satisfies three properties: *local accuracy* (the sum of feature contributions equals the model prediction), *missingness* (features not used in a prediction receive zero attribution), and *consistency* (if a model changes so that a feature's contribution increases, its attribution does not decrease) [21].

SHAP can be applied to any model, including text classifiers, by approximating contributions using methods such as Kernel SHAP. In applications like fraud detection, SHAP helps identify which tokens or terms influenced the model's decision, allowing for verification and pattern analysis. This makes it possible to relate specific linguistic elements to model outputs without altering an underlying classifier [21].

III. Dataset

To support both the training of the proposed fraud detection model and the evaluation of its performance in a real-world context, two distinct datasets were constructed. The first dataset was used to develop and validate the machine learning classifier, incorporating a diverse set of manually annotated examples collected from multiple sources. The second dataset was designed to monitor ad activity over an extended period, enabling the classifier to operate in a live recovery pipeline and supporting manual verification for evidence collection. Together, these datasets provide the empirical foundation for both model training and investigative analysis of fraudulent campaigns on Meta Ads.

A. Dataset for Model Development

The training and validation of the machine learning classifier were based on a heterogeneous dataset composed of seven subsets, collected using both automated and manual methods. These subsets were designed to capture a broad spectrum of fraudulent scenarios and legitimate ad content on Meta-platforms. Data were acquired via the Meta Ads API (limited to ads classified as "sensitive"), as well as through custom web scraping and manual collection strategies.

Ads Referencing Broadcast TV Companies This subset includes deceptive advertisements collected in April and July 2024, which exploit public trust by mimicking the identity of major broadcast television outlets. These ads often refer to well-known programs, public figures such as journalists

or actors, and recognizable visual elements to transmit legitimacy and increase the likelihood of user engagement.

Extreme Climate Events Publications that reference the humanitarian and social context of the 2024 floods in southern Brazil. This collection includes both legitimate appeals and fraudulent messages seeking to exploit public solidarity.

Government programs Advertising related to official financial and travel initiatives, such as national debt relief and subsidized air travel. These campaigns are frequent targets of impersonation and misappropriation by fraudsters seeking to exploit their institutional legitimacy.

State-linked Content General references to government-affiliated communication used in deceptive contexts. This category includes both sensitive and non-sensitive advertisements and primarily comprises fraudulent content.

Quantitatively, the consolidated dataset contains a total of 3,365 unique advertisements, with a notable class imbalance: 72.5% were labeled as fraudulent and 27.5% as non-fraudulent.

The dataset varies in terms of metadata richness. API-based data includes audience targeting and impression estimates, whereas scraped and manually collected ads are limited to textual content and visual elements. For consistency, only the textual component was used to train the model. Duplicate messages, common in segmented ad variants, were removed to avoid skewed results or overfitting.

B. Dataset for Monitoring and Evidence Analysis

A second dataset was assembled to enable the longitudinal assessment of the model's performance and to guide the targeted manual annotation of advertisements for investigative purposes.

The monitoring period spanned from July 1st to November 27th, 2024, during which 41,521 ad entries were collected. The trained classifier was applied to each collected ad, assigning a probability score of fraud. Advertisements with a score greater than or equal to 0.5 were flagged as potentially fraudulent and prioritized for manual review. This decision threshold was chosen to favor recall, enabling the capture of a broader set of suspicious ads for subsequent human evaluation.

Figure 1 illustrates the operational pipeline supporting this process. Ads were initially collected using a custom scraper that queried the Meta Ads Library. These raw entries were then processed through a data enrichment step, in which relevant metadata was structured and indexed. The classifier's inference scores were computed based on features extracted from the textual content of each ad. Scored ads were passed through a relevance filter that retained only recent entries (up to one week old) with scores above the defined threshold. These filtered ads were

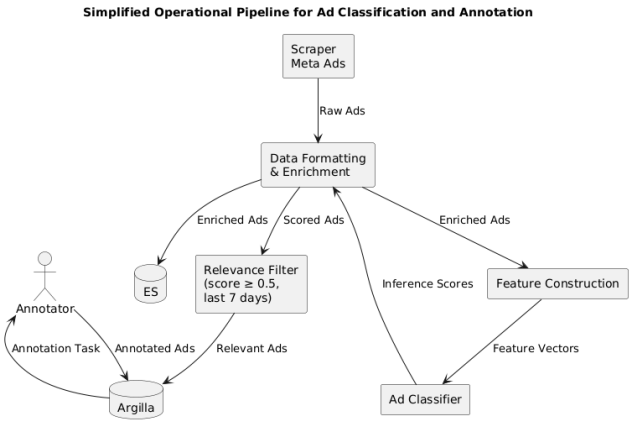

Fig. 1. Operational pipeline for fraud detection in Meta Ads. Ads are collected, enriched, classified by a trained model, filtered by score and recency, and then reviewed and labeled by human annotators using Argilla.

then published to an annotation platform (Argilla) [19], where annotators verified the content and classified the ad as fraudulent or legitimate. Final annotated data was used both for evidence gathering and future model refinements.

The 10,826 unique advertisements flagged as fraudulent by the classifier underwent manual review. Each ad was annotated to confirm whether it constituted fraud and, if applicable, to specify the corresponding fraud category. The analyses presented in this study are based exclusively on these manually labeled samples, ensuring the reliability of the findings and supporting the interpretability of the identified fraud patterns.

IV. Training and Validation Protocol

The classifier developed in this study was trained using SVM with a RBF kernel. This architecture was chosen because of its well-documented effectiveness in handling high-dimensional, sparse feature spaces, such as those generated through TF-IDF vectorization of short text.

To ensure reliable model selection and generalization, we adopted a nested 10-fold cross-validation strategy. The inner loop was used for hyperparameter tuning, systematically evaluating configurations of the RBF-SVM. We explored two key parameters: the regularization constant C, which mediates the trade-off between maximizing the margin and minimizing the classification error, and the kernel coefficient γ, which controls the influence range of individual training samples in the feature space. Higher values of C penalize misclassifications more heavily, potentially narrowing the decision boundary, while lower values of γ produce smoother, more generalized models. A grid search was conducted on $C \in \{1, 10, 100\}$ and $\gamma \in \{0.01, 0.1\}$, allowing a controlled exploration of the complexity of the model and the generalization potential.

The outer loop also employed 10-fold cross-validation to assess generalization performance. Stratification was performed over the joint distribution of the class label and

the data source to preserve the representational diversity of the dataset, balancing fraudulent and non-fraudulent examples in manual collections, API-based, and scraped collections.

Model performance was evaluated using a comprehensive set of metrics:

ROC and ROC AUC
: These metrics assess the model's discrimination capability across all decision thresholds. Although ROC (Receiver Operating Characteristics) curves visualize the trade-off between true positive and false positive rates, the area under the curve (AUC) provides an aggregate measure of separability.

F1 and F2 scores
: F1 is the harmonic mean of precision and recall, providing a balanced evaluation. F2, in contrast, places more weight on recall, a strategic choice in our setting. Since flagged ads underwent manual verification, our priority was to minimize the number of undetected fraudulent cases, even at the cost of increasing false positives.

Recall and Precision
: Recall quantifies the model's ability to retrieve actual fraudulent ads, while precision measures the accuracy of fraud predictions between all flagged items, both crucial for operating fraud detection workflows.

To further enhance the interpretability of the model decisions, we applied the SHAP (SHapley Additive Explanations) analysis. This post hoc method attributes the classifier's output to individual features, allowing us to quantify the contribution of specific terms to each prediction. The SHAP analysis revealed which keywords and linguistic patterns were most influential in fraud detection, providing transparency into the internal logic of the model and valuable insights into the rhetorical strategies employed in deceptive ads.

V. Results

The performance of the trained classifier was evaluated using the outer loop of the nested cross-validation, employing a 10-fold partitioning scheme stratified by class label and data source. The results, summarized in Table I, report the mean and standard deviation for the five main evaluation metrics across all folds: F1 score, F2 score, recall, precision, and ROC AUC.

The classifier achieved a mean F1 score of 0.91 ($\pm$ 0.01) and a mean F2 score of 0.93 ($\pm$ 0.01), indicating a strong balance between precision and recall, with an emphasis on sensitivity. The average recall was 0.95 ($\pm$ 0.02), demonstrating the ability of the model to identify the majority of fraudulent advertisements. Although precision was slightly lower, at 0.88 ($\pm$ 0.02), this reflects an intentional trade-off that favors recall, consistent with the study's objective of minimizing false negatives in a context where manual validation is available. The mean ROC AUC of 0.88 ($\pm$ 0.02) confirms the reliability of the model in distinguishing between fraudulent and legitimate advertisements at decision thresholds.

TABLE I
Mean and Standard Deviation across outer loop cross-validation folds.

Metric	Mean	Std. Dev.
F1 Score	0.91	0.01
F2 Score	0.93	0.01
Recall	0.95	0.02
Precision	0.88	0.02
ROC AUC	0.88	0.02

The discriminatory capacity of the model is further illustrated in the ROC curve shown in Figure 2. The figure displays the ROC curves obtained from each of the 10 cross-validation folds as dotted lines, alongside the averaged curve (in blue), which aggregates classifier performance across all folds. The area under the mean ROC curve (AUC) is 0.88 ± 0.02, indicating a consistently strong ability to distinguish between fraudulent and legitimate ad in various threshold settings.

The shaded region around the mean ROC curve represents one standard deviation, capturing the variability in the true positive rate multiple times for each corresponding false positive rate. This provides a visual measure of the robustness of the model. In addition, a dashed red diagonal line represents the performance of a random classifier (i.e., chance level), serving as a baseline reference.

In the figure legend, the average F1 and F2 scores are also presented (0.91 and 0.93, respectively, each with a standard deviation of 0.01), reinforcing the balance- and recall-oriented performance profile of the model. Together, these results confirm that the classifier maintains reliable sensitivity and specificity even under variation in the training data partitions.

A. SHAP-Based Model Interpretation

To interpret the internal logic of the trained model and better understand which textual features influenced its decisions, we applied SHAP analysis to the classifier. The results have shown a mix of terms commonly found in *fraudulent advertising rhetoric*, such as imperatives and call-to-action expressions ("click", "watch", "talk", "learn", "send"), as well as terms related to urgency or exclusivity ("free", "solution", "need", "better"). These lexical patterns suggest that the classifier learned to associate persuasive marketing language, often exaggerated or manipulative, with fraudulent behavior.

Furthermore, the presence of generic placeholders such as "name", "brand", and "#NUMBER" reflects common structures in templated scam messages, where dynamic content is automatically injected to personalize or regionalize the ad. The appearance of specific program names (e.g., "desenrola") and institutional references (e.g. "brazil", "program") indicates that the model also de-

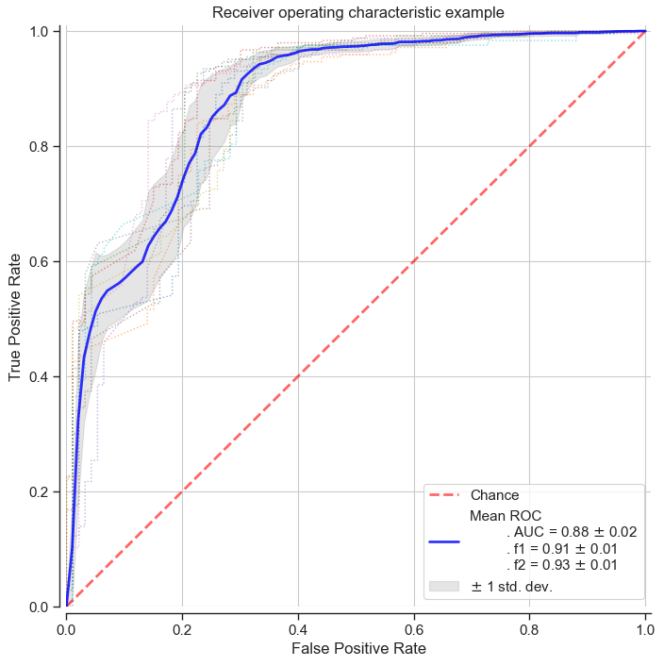

Fig. 2. ROC curves for all cross-validation folds (dotted), with mean ROC curve (solid blue), standard deviation band (gray), and random classifier baseline (dashed red).

tected the frequent misuse of government or social initiatives to transmit credibility to fraudulent campaigns.

Terms referencing *products and platforms*, such as "product", "link", "instagram", "whatsapp" and "vivo" (a major telecom company in Brazil), reveal that the classifier was sensitive to the frequent linkage of fraud to known brands or communication tools, either to simulate legitimacy or facilitate contact through messaging apps. In particular, terms related to medical and wellness such as "health", "life", and "treatment" also appeared of high importance, aligning with the observation that many scam-ads exploit health-related anxieties to deceive users.

Interestingly, even emoji tokens (converted into textual labels during preprocessing, e.g., "checkmarkbutton") contributed meaningfully to classification. These symbols often serve to visually highlight "benefits" or create a false sense of support and trust.

Overall, the SHAP analysis confirms that the model effectively internalized patterns aligned with known fraud strategies: misuse of institutional language, aggressive sales rhetoric, brand impersonation, and emotionally charged messaging. These insights validate the relevance of the feature engineering strategy and provide transparency in the decision-making process of the model.

VI. Longitudinal Analysis of the Classifier Operation

Following model deployment, the classifier was applied in a real-world monitoring context using the pipeline described in Figure 1. Between July and November 2024, a total of 10,826 advertisements were automatically flagged as potentially fraudulent and later manually reviewed by human annotators.

The classifier achieved an overall precision of 77% in this manually annotated subset. Among the ads predicted as fraudulent (that is, those with a model score greater than or equal to 0.5), 8,287 were confirmed as fraud by the annotators, while 2,539 were labeled as legitimate. This corresponds to a false positive rate of approximately 23%, which is consistent with the intentional bias of the model toward recall.

It is important to emphasize that this analysis was restricted to the subset of ads with high predicted probability scores (≥ 0.5), as defined by the relevance filter. This threshold reflects the model's sensitivity-oriented calibration. The goal was not to minimize false positives but to ensure that truly fraudulent cases were rarely missed. Given that all flagged ads underwent human validation, the operational focus was on maximizing the inclusion of suspicious content for expert verification, rather than filtering too aggressively at the model level.

This operational analysis confirms that the classifier performs reliably when integrated into a human-in-the-loop annotation workflow. Moreover, the false positives, while present, remained within acceptable limits for investigative contexts where the cost of omission is higher than that of over-inclusion.

A. Coverage–Precision Trade-off

To further investigate the model's decision behavior and the trade-off between coverage and precision, we performed an analysis of the results of manual annotation segmented by classifier score thresholds. Figure 3 presents the distribution of ads, both fraudulent and legitimate, as a function of score cutoffs ranging from 0.50 to 1.00.

As expected, the proportion of confirmed fraud cases increases as the threshold increases. At the default operational threshold of 0.5, the model achieves high recall but includes a significant portion of false positives. However, as the threshold increases, the volume of ads decreases, while the relative concentration of true positives improves substantially. For example, above 0.85, the vast majority of ads are confirmed frauds.

This progressive filtering effect illustrates a clear confidence gradient in model scoring. Although a lower threshold ensures broader fraud detection (beneficial when manual validation is available), higher thresholds can be used to prioritize review efforts or automate downstream decisions with greater precision. These insights may inform future deployment scenarios in which the annotation workload must be dynamically adjusted based on capacity or risk tolerance.

B. Lifespan Analysis of Legitimate and Fraudulent Ads

We also examine the duration of the activity of the ads, comparing the typical lifetime of fraudulent and legitimate

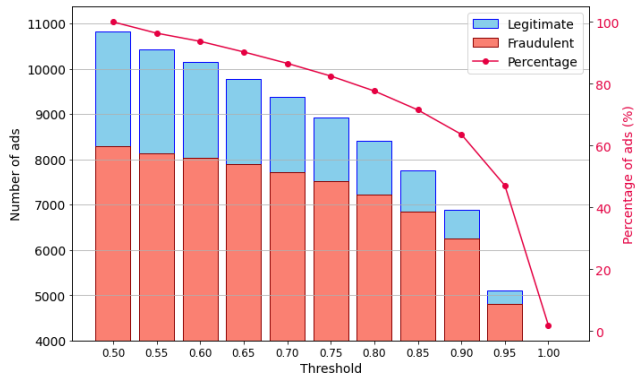

Fig. 3. Distribution of annotated ads by score threshold. Bars indicate the number of ads classified as fraudulent or legitimate per threshold bin. The red line (right axis) shows the cumulative percentage of all annotated ads included at each cutoff.

ads. The results indicate a clear behavioral distinction between the two classes. Legitimate ads tend to remain active for longer periods, suggesting their integration into more stable and ongoing advertising campaigns. In particular, a significant portion of legitimate ads (433 instances) persisted on the platform for more than 17 days.

In contrast, fraudulent ads exhibit much shorter lifespans. The vast majority were active for only one or two days (1,967 and 2,716 cases respectively). This pattern suggests an intentional evasion strategy in which malicious actors limit the exposure window of their campaigns to reduce the chances of detection or removal by the platform's moderation mechanisms. This short-lived nature also reinforces the importance of frequent monitoring and rapid classification cycles, as fraudulent content can be ephemeral and highly dynamic.

These behavioral differences support the notion that temporal features, such as ad persistence, serve as complementary signals for fraud detection and highlight the necessity of proactive and responsive infrastructure for near real-time monitoring.

C. Narratives of Deception: Impersonation and Thematic Diversity

A qualitative review of the annotated fraudulent ads reveals a recurring strategy of impersonating well-known Brazilian television personalities to promote scams. Figure 4 presents the distribution of fraud categories between different individuals whose names or likenesses were misused in deceptive campaigns.

The most frequently targeted figure was the physician and TV host *Drauzio Varella*, associated with more than 7,400 fraudulent ads. These campaigns covered a wide range of topics, including sexual health products, rejuvenation treatments, diabetes cures, and financial fraud. Other high-profile personalities, such as *Ana Maria Braga*, *William Bonner* and *Fátima Bernardes*, were also impersonated in hundreds of scam ads.

This pattern highlights the strategic use of public trust and media credibility as tools of deception. Scammers exploit the authority and familiarity of these individuals to legitimize dubious claims, particularly in advertisements related to health, wellness, and financial advice. The diversity of fraud types, from chronic disease cures to anti-aging products, and even hearing or vision treatments, illustrates the breadth of narratives employed to attract attention and provoke engagement.

Such abuse of public figures not only poses risks to consumers, but also damages the reputation of those impersonated. It underscores the urgency for platform-level interventions capable of detecting both semantic deception and unauthorized use of identities in digital advertising.

VII. Conclusion

This study presented a machine learning approach to detect fraudulent advertisements on the Meta Ads platform, focusing exclusively on textual content. Through the construction of a manually annotated dataset and the application of natural language processing techniques, we trained a SVM classifier capable of distinguishing fraudulent content with high recall and consistent operational performance.

The classifier demonstrated strong generalization in cross-validation and maintained reliable behavior when deployed in an acquisition-time annotation pipeline. Empirical analysis showed that the model effectively prioritized suspicious ads for human review, capturing a wide spectrum of fraud campaigns. The model achieved a mean F2 score of 0.93 ± 0.01, reflecting its high sensitivity and suitability to minimize false negatives. Qualitative investigations revealed that these scams often exploit the image of trusted public figures and adopt emotionally charged narratives related to health, finances, and personal well-being. Temporal analyzes also indicated distinct behavioral patterns, with fraudulent ads typically exhibiting short lifespans compared to more stable legitimate campaigns.

Although the current model focuses solely on text, future developments will seek to incorporate additional feature sources. One promising direction is the inclusion of visual information extracted from ad images, which may carry persuasive cues (e.g., branding, facial impersonation, product visuals) crucial for modeling deception. Moreover, metadata such as ad longevity and recurrence patterns can be integrated into the model architecture or used to refine detection post-deployment.

These extensions aim to enhance the model's sensitivity and robustness without sacrificing interpretability. Ultimately, the development of multimodal and temporally-aware classifiers may offer more comprehensive protection against digital ad fraud in real-world monitoring environments.

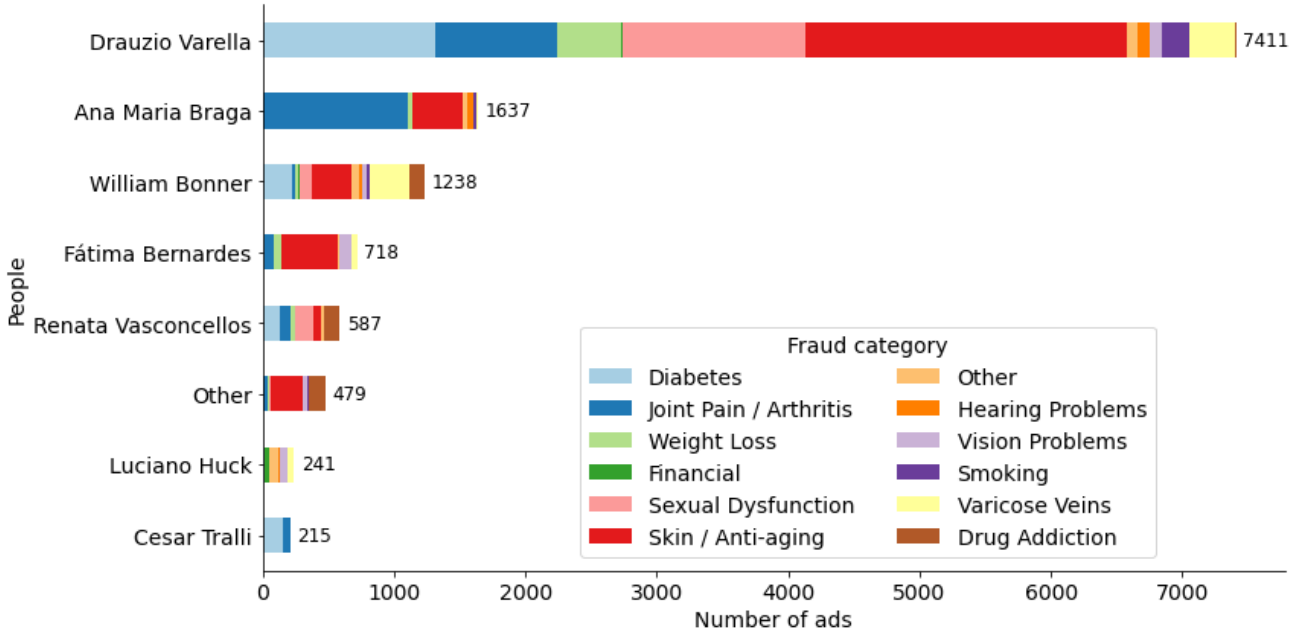

Fig. 4. Distribution of fraud categories by impersonated public figures. Each color-coded bar segment represents a specific type of fraud campaign associated with the named individual.

References

[1] S. Sadeghpour and N. Vlajic, "Click Fraud in Digital Advertising: A Comprehensive Survey," in *Proc. ACM Conference on Computer and Communications Security*, vol. 4, pp. 765–776, 2013.

[2] J.-L. Richet, "How cybercriminal communities grow and change: An investigation of ad-fraud communities," *Technological Forecasting and Social Change*, vol. 174, p. 121282, 2022.

[3] A. Kindi, M. S. Islam, and N. Rahman, "AI-Driven Fraud Detections in Financial Institutions: A Comprehensive Study," *Journal of Computer Science and Technology Studies*, vol. 7, 2025. doi: 10.32996/jcsts.2025.7.1.8.

[4] R. M. Santini *et al.*, "Golpe financeiro através de anúncios no Meta Ads," NetLab UFRJ, Relatório Técnico, abr. 2023. [Online]. Available: https://www.netlab.eco.ufrj.br

[5] R. M. Santini, D. Salles, B. M. Martins, A. Moreira, and J. G. Haddad, "Seeing through opacity: The limitations of digital ad transparency

[6] S. Sadeghpour and N. Vlajic, "Ads and fraud: A comprehensive survey of fraud in online advertising," *Journal of Cybersecurity and Privacy*, vol. 1, no. 4, pp. 804–832, 2021.

[7] S. Bharne and P. Bhaladhare, "Comprehensive Analysis of Online Social Network Frauds," in *Proc. Int. Conf. on Advances in Data-driven Computing and Intelligent Systems*, pp. 23–40, 2022. Springer.

[8] D. Dasgupta, Z. Akhtar, and S. Sen, "Machine learning in cybersecurity: a comprehensive survey," *The Journal of Defense Modeling and Simulation*, vol. 19, no. 1, pp. 57–106, 2022.

[9] J. Wells, "Meta Saw Signs of Fraud on Its Platforms. It Took a Year to Act.," *The Wall Street Journal*, news article, May. 28, 2025. [Online]. Available: https://www.wsj.com/tech/meta-fraud-facebook-instagram-813363c8

[10] T. Matschak, S. Trang, and C. Prinz, "A taxonomy of machine learning-based fraud detection systems," in *Proc. European Conference on Information Systems (ECIS)*, 2022.

[11] D. Sisodia and D. S. Sisodia, "Gradient boosting learning for fraudulent publisher detection in online advertising," *Data Technologies and Applications*, vol. 55, no. 2, pp. 216–232, 2021.

[12] A. M. Qatawneh, "The role of artificial intelligence in auditing and fraud detection in accounting information systems: moderating role of natural language processing," *International Journal of Organizational Analysis*, 2024.

[13] P. Boulieris, J. Pavlopoulos, A. Xenos, and V. Vassalos, "Fraud detection with natural language processing," *Machine Learning*, vol. 113, no. 8, pp. 5087–5108, 2024.

[14] V. R. Saddi, B. Gnanapa, S. Boddu, and J. Logeshwaran, "The Role of Natural Language Processing in Detecting Insurance Fraud," in *Proc. 4th Int. Conf. on Communication, Computing and Industry 6.0 (C216)*, pp. 1–6, 2023. IEEE.

[15] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[16] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[17] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, Article 31, pp. 1–41, 2022. doi: 10.1145/3495162.

[18] V. Dogra, S. Verma, Kavita, P. Chatterjee, J. Shafi, J. Choi, and M. F. Ijaz, "A complete process of text classification system using state-of-the-art NLP models," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 1883698, 2022.

[19] Argilla, "Argilla: Open-source data labeling and curation for NLP," 2025. [Online]. Available: https://argilla.io/

[20] A. Al-Khafaji and O. Karan, "Explainable AI for Predicting User Behavior in Digital Advertising," in *Emerging Trends and Applications in Artificial Intelligence*, F. P. García Márquez, A. Jamil, A. A. Hameed, and I. Segovia Ramírez, Eds., Cham: Springer, pp. 520–531, 2024.

[21] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.